

Multiple–Instance Learning: Radon–Nikodym Approach to Distribution Regression Problem.

Vladislav Gennadievich Malyshkin*

Ioffe Institute, Politekhnikeskaya 26, St Petersburg, 194021, Russia

(Dated: November, 27, 2015)

\$Id: DistReg1Step.tex,v 1.41 2015/12/02 11:00:50 mal Exp \$

For distribution regression problem, where a bag of x –observations is mapped to a single y value, a one–step solution is proposed. The problem of random distribution to random value is transformed to random vector to random value by taking distribution moments of x observations in a bag as random vector. Then Radon–Nikodym or least squares theory can be applied, what give $y(x)$ estimator. The probability distribution of y is also obtained, what requires solving generalized eigenvalues problem, matrix spectrum (not depending on x) give possible y outcomes and depending on x probabilities of outcomes can be obtained by projecting the distribution with fixed x value (delta–function) to corresponding eigenvector. A library providing numerically stable polynomial basis for these calculations is available, what make the proposed approach practical.

* malyshki@ton.ioffe.ru

I. INTRODUCTION

Multiple instance learning[1] is an important Machine Learning (ML) concept having numerous applications[2]. In multiple instance learning class label is associated not with a single observation, but with a “bag” of observations. A very close problem is distribution regression problem, where a sample distribution of x is mapped to a single y value. There are numerous heuristics methods developed from both: ML and distribution regression sides, see [3, 4] for review.

As in any ML problem the most important part is not so much the learning algorithm, but the way how the learned knowledge is represented. Learned knowledge is often represented as a set of propositional rules, regression function, Neural Network weights, etc. In this paper we consider the case where knowledge is represented as a function of distribution moments. Recent progress in numerical stability of high order moments calculation[5] allow the moments of very high order to be calculated, e.g. in Ref. [6] up to hundreds, thus make this approach practical.

Most of distribution regression algorithms deploy a two-step type of algorithm[4] to solve the problem. In our previous work [7] a two-step solution with knowledge representation in a form of Christoffel function was developed. However, there is exist a one-step solution to distribution regression problem, a random distribution to random value, that converts each bag’s observations to moments of it, then solving the problem random vector (the moments of random distribution) to random value. Once this transition is made an answer of least squares or Radon–Nikodym type from Ref. [5] can be applied and close form result obtained. The distribution of outcomes, if required, can be obtained by solving generalized eigenvalues problem, then matrix spectrum give possible y outcomes, and the square of projection of localized at given x bag distribution to eigenvector give each outcome probability. This matrix spectrum ideology is similar to the one we used in [7], but is more generic and not reducible to Gauss quadrature.

The paper is organized as following: In Section II a general theory of distribution regression is discussed and close form result or least squares and Radon–Nikodym type are presented. Then in Section III an algorithm is described and numerical example of calcula-

tions is presented. In Section IV possible further development is discussed.

II. ONE-STEP SOLUTION

Consider distribution regression problem where a bag of N observations of x is mapped to a single outcome observation y for $l = [1..M]$.

$$(x_1, x_2, \dots, x_j, \dots, x_N)^{(l)} \rightarrow y^{(l)} \quad (1)$$

A distribution regression problem can have a goal to estimate y , average of y , distribution of y , etc. given specific value of x

For further development we need x basis $Q_k(x)$ and some x and y measure. For simplicity, not reducing the generality of the approach, we are going to assume that x measure is a sum over j index $\sum_{j=1}^N$, y measure is a $\sum_{l=1}^M$, the basis functions $Q_k(x)$ are polynomials $k = 0..d_x - 1$, where d_x is the number of elements in x basis, typical value for d_x is below 10–15.

Let us convert the problem “random distribution” to “random variable” to the problem “vector of random variables” to “random variable”. The simplest way to obtain “vector of random variables” from $x_j^{(l)}$ distributions is to take the moments of it. Now the $\langle Q_k \rangle^{(l)}$ would be this random vector:

$$\langle Q_k \rangle^{(l)} = \sum_{j=1}^N Q_k(x_j^{(l)}) \quad (2)$$

$$\left(\langle Q_0 \rangle^{(l)}, \dots, \langle Q_{d_x-1} \rangle^{(l)} \right) \rightarrow y^{(l)} \quad (3)$$

Then the (3) becomes vector to value problem. Introduce

$$Y_q = \sum_{l=1}^M y^{(l)} \langle Q_q \rangle^{(l)} \quad (4)$$

$$(G)_{qr} = \sum_{l=1}^M \langle Q_q \rangle^{(l)} \langle Q_r \rangle^{(l)} \quad (5)$$

$$(yG)_{qr} = \sum_{l=1}^M y^{(l)} \langle Q_q \rangle^{(l)} \langle Q_r \rangle^{(l)} \quad (6)$$

The problem now is to estimate y (or distribution of y) given x distribution, now mapped to a vector of moments $\langle Q_k \rangle$ calculated on this x distribution. Let us denote these input

moments as M_k to avoid confusion with measures on x and y . For the case we study the x value is given, and for a state with exact x the M_k values are:

$$M_k(x) = NQ_k(x) \quad (7)$$

what means that all N observations in a bag give exactly the same x value. The problem now becomes a standard: random vector to random variable. We have solutions of two types for this problem, see [5] Appendix D, Least Squares A_{LS} and Radon–Nikodym A_{RN} . The answers would be:

$$A_{LS}(x) = \sum_{q,r=0}^{d_x-1} M_q(x) (G)_{qr}^{-1} Y_r \quad (8)$$

$$A_{RN}(x) = \frac{\sum_{q,r,s,t=0}^{d_x-1} M_q(x) (G)_{qr}^{-1} (yG)_{rs} (G)_{st}^{-1} M_t(x)}{\sum_{q,r=0}^{d_x-1} M_q(x) (G)_{qr}^{-1} M_r(x)} \quad (9)$$

The (8) is least squares answer to y estimation given x . The (9) is Radon–Nikodym answer to y estimation given x . These are the two y estimators at given x for distribution regression problem 1. These answers can be considered as an extension of least squares and Radon–Nikodym type of interpolation from value to value problem to random distribution to random variable problem. In case $N = 1$ the A_{LS} and A_{RN} are reduced exactly to value to value problem considered in Ref. [5]. Note, that the $A_{LS}(x)$ answer not necessary preserve y sign, but $A_{RN}(x)$ always preserve y sign, same as in value to value problem.

If y distribution at given x need to be estimated this problem can also be solved. With one–step approach of this paper we do not need $Q_m(y)$ basis used in two–step approach of Ref. [7] and outcomes of y are estimated from x moments only. Generalized eigenvalues problem[5] give the answer:

$$\sum_{r=0}^{d_x-1} (yG)_{qr} \psi_r^{(i)} = y^{(i)} \sum_{r=0}^{d_x-1} (G)_{qr} \psi_r^{(i)} \quad (10)$$

The result of (10) is eigenvalues $y^{(i)}$ (possible outcomes) and eigenvectors $\psi^{(i)}$ (can be used to compute the probabilities of outcomes). The problem now becomes: given x value estimate possible y –outcomes and their probabilities. The moments of states with given x value are $NQ_q(x)$ from (7), so the distribution with (7) moments should be projected to distributions corresponding to $\psi_q^{(i)}$ states, the square of this projection give the weight and normalized weight give the probability. This is actually very similar to ideology we used in [7], but the

eigenvalues from (10) no longer have a meaning of Gauss quadrature nodes. The eigenvectors $\psi_r^{(i)}$ correspond to distribution with moments $\langle Q_q \rangle = \sum_{r=0}^{d_x-1} (G)_{qr} \psi_r^{(i)}$, and the distribution with such moments correspond to $y^{(i)}$ value. These distributions can be considered as “natural distribution basis”. This is an important generalization of Refs. [5, 6] approach to random distribution, where natural basis for random value, not random distribution, was considered.

The projection of two x distributions with moments $M_k^{(1)}$ and $M_k^{(2)}$ on each other is

$$\langle M^{(1)} | M^{(2)} \rangle_\pi = \sum_{q,r=0}^{d_x-1} M_q^{(1)} (G)_{qr}^{-1} M_r^{(2)} \quad (11)$$

then the required probabilities, calculated by projecting the (7) distribution to natural basis states, are:

$$w^{(i)}(x) = \left(\sum_{r=0}^{d_x-1} M_r(x) \psi_r^{(i)} \right)^2 \quad (12)$$

$$P^{(i)}(x) = w^{(i)}(x) / \sum_{r=0}^{d_x-1} w^{(r)}(x) \quad (13)$$

The (10) and (13) is one-step answer to distribution regression problem: find the outcomes $y^{(i)}$ and their probabilities $P^{(i)}(x)$. Note, that in this setup possible outcomes $y^{(i)}$ do not depend on x , and only probabilities $P^{(i)}(x)$ of outcomes depend on x . This is different from a two-step solution of [7] where outcomes and their probabilities both depend on x . Also note that $\sum_{r=0}^{d_x-1} w^{(r)}(x) = \sum_{q,r=0}^{d_x-1} M_q(x) (G)_{qr}^{-1} M_r(x)$.

One of the major difference between the probabilities (13) and probabilities from Christoffel function approach [7] is that the (13) has a meaning of “true” probability while in two-step solution [7] Christoffel function value is used as a proxy to probability on first step. It is important to note how the knowledge is represented in these models. The model (8) has learned knowledge represented in d_x by d_x matrix (5) and d_x size vector (4). The model (9) as well as distribution answer (13) has learned knowledge represented in two d_x by d_x matrices (5) and (6).

III. NUMERICAL ESTIMATION OF ONE-STEP SOLUTION

Numerical instability similar to the one of two-stage Christoffel function approach [7] also arise for approach in study, but now the situation is much less problematic, because we

do not have y -basis $Q_m(y)$, and all the dependence on y enter the answer through matrix (6). In this case the only stable x basis $Q_k(x)$ is required.

The algorithm for y estimators of (8) or (9) is this: Calculate $\langle Q_k \rangle^{(l)}$ moments from (2, then calculate matrices (5) and (6), if least squares approximation is required also calculate moments (4). In contrast with Christoffel function approach where $\langle Q_q Q_r \rangle$; $q, r = [0..d_x - 1]$ matrix can be obtained from Q_k ; $k = [0..2d_x - 1]$ moments by application of polynomials multiplication operator, here the (5) and (6) can be hardly obtained this way for $N > 1$ and should be calculated directly from sample. This is not a big issue, because d_x is typically not large. Then inverse matrix $(G)_{qr}$ from (5), this matrix is some kind similar to Gramm matrix, but uses distribution moments, not basis functions. Finally put all these to (8) for least squares $y(x)$ estimation or to (9) for Radon–Nikodym $y(x)$ estimation.

If y -distribution is required then solve generalized eigenvalues problem (10), obtain $y^{(i)}$ as possible y -outcomes (they do not depend on x), and calculate x -dependent probabilities (13), these are squared projection coefficient of a state with specific x value, point-distribution (7), or some other x distribution of general form, to $\psi^{(i)}$ eigenvector.

To show an application of this approach let us take several simple distribution to apply the theory. Let ϵ be a uniformly distributed $[-1; 1]$ random variable and take $N = 1000$ and $M = 10000$. Then consider sample distributions build as following 1) For $l = [1..M]$ take random x out of $[-1; 1]$ interval. 2) Calculate $y = f(x)$, take this y as $y^{(l)}$. 3) Build a bag of x observations as $x_j = x + R\epsilon$; $j = [1..N]$, where R is a parameter. The following three $f(x)$ functions for building sample distribution are used:

$$f(x) = x \tag{14}$$

$$f(x) = \frac{1}{1 + 25x^2} \tag{15}$$

$$f(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases} \tag{16}$$

In Figs. 1, 2, 3, the (8) and (9) answers are presented for $f(x)$ from (14), (15) and (16) respectively for $R = \{0.1, 0.3\}$ and $d_x = \{10, 20\}$. The x range is specially taken slightly wider than $[-1; 1]$ interval to see possible divergence outside of measure support. In most cases Radon–Nikodym answer is superior, and in addition to that it preserves the sign of y . Least squares approximation is good for special case $f(x) = x$ and typically diverges at x outside of measure support.

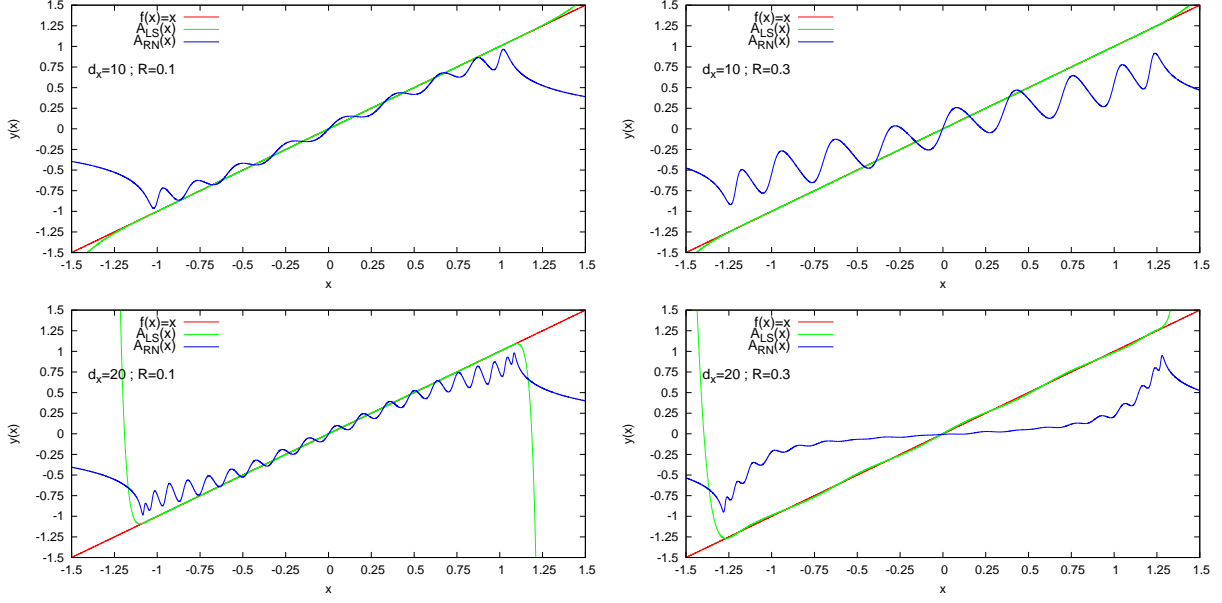


FIG. 1. The $y(x)$ estimation for $f(x)$ from (14).

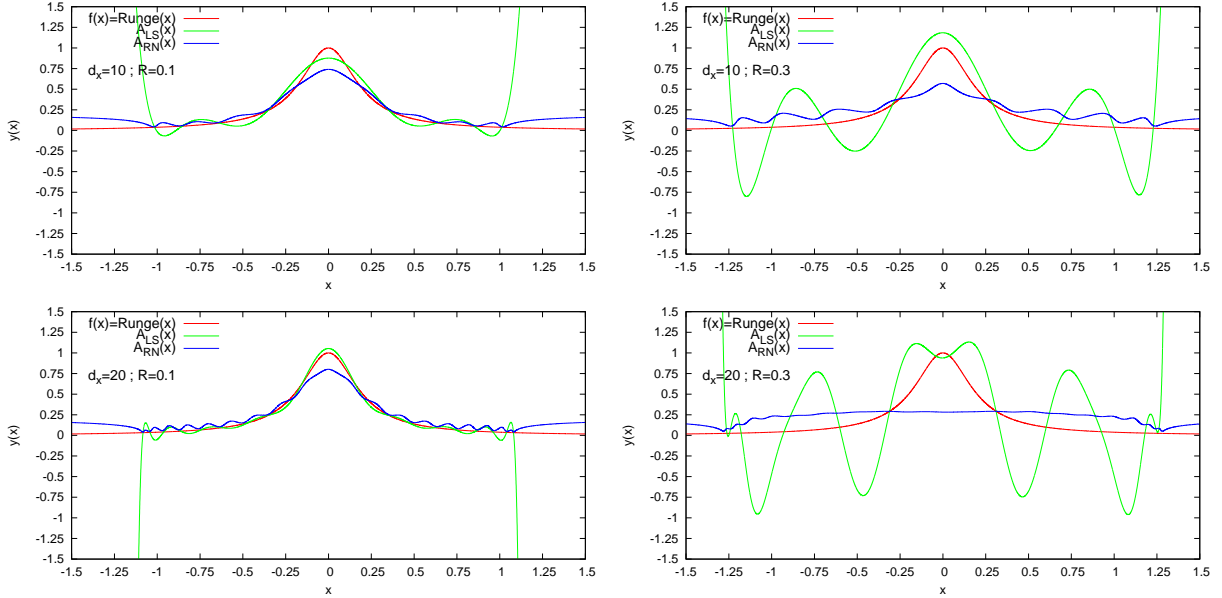


FIG. 2. The $y(x)$ estimation for $f(x)$ from (15).

The numerical estimation of probability function (the $y^{(i)}$ and $P^{(i)}(x)$) were also calculated and eigenvalue index i , corresponding to maximal P typically correspond to $y^{(i)}$, for which $f(x)$ is most close. For simplistic case (14) see Fig. 4. See the Ref. [8], file com/polytechnik/algorithms/ ExampleDistribution1Stage.scala for algorithm implementation.

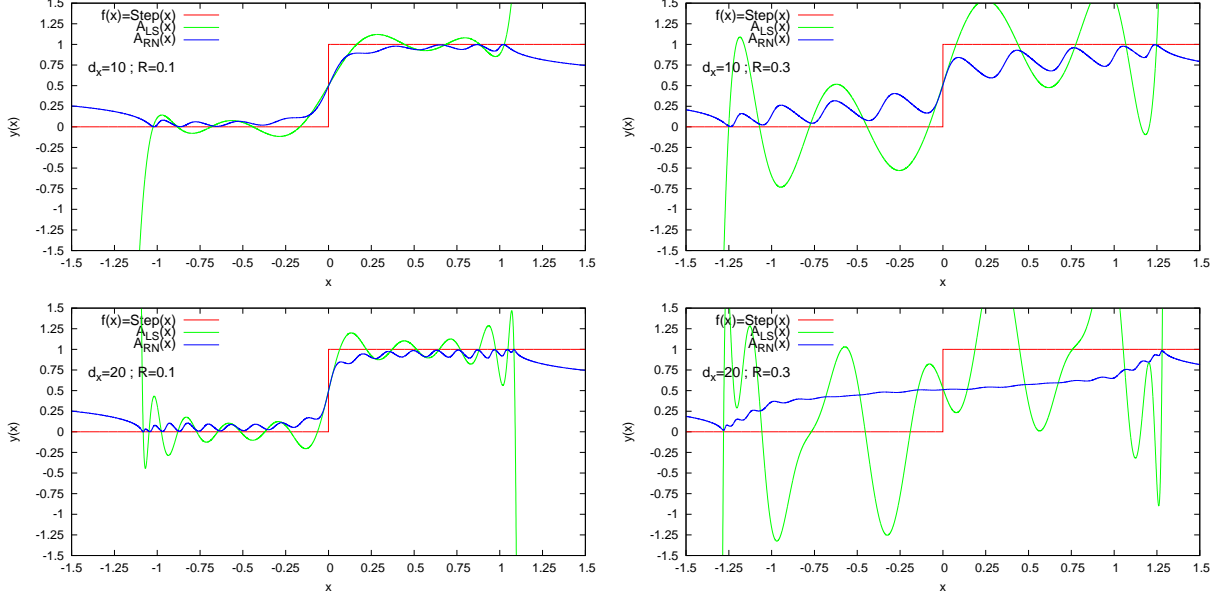


FIG. 3. The $y(x)$ estimation for $f(x)$ from (16).

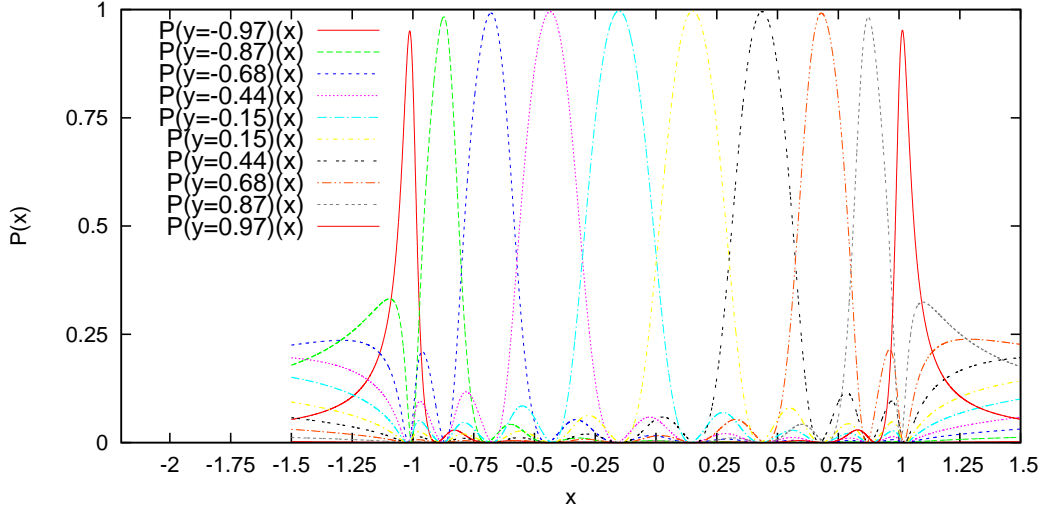


FIG. 4. Probabilities for all $d_x = 10$ outcomes of $y^{(i)}$ as a function of x for $f(x)$ from (14).

IV. DISCUSSION

In this work a one-stage approach is applied to distribution regression problem. The bag's observations are initially converted to moments, then least squares or Radon–Nikodym theory can be applied and closed form answer to be received. These (8) and (9) estimate y value given x . This answer can be generalized to “what is y estimate given distribution of x ”. For this problem obtain moments $\langle Q_k \rangle$, corresponding to given distribution of x ,

first, then use them in (8) or (9) instead of $M_k(x)$, corresponding to localized at x state. Similarly, if probabilities of y outcomes are required for given distribution of x , the $\langle Q_k \rangle$ should be used in weights expression (12) instead of $M_k(x)$ (this is a special case of two distribution projection on each other (11)). Computer code implementing the algorithms is available[8].

And in conclusion we want to discuss possible directions of future development.

- In this work a closed form solution for random distribution to random value problem (1) is found. The question arise about problem order increase, replace “random distribution” by “random distribution of random distribution” (or even further “random distribution of random distribution of random distribution”, etc.). In this case each x_j in (1) should be treated as a sample distribution itself, and index j then can be treated as 2D index x_{j_1, j_2} . Working with 2D indexes is actually very similar to working with images, see Ref. [6] where the 2D index was used for image reconstruction by applying Radon–Nikodym or least squares approximation. Similarly, the results of this paper, can be generalized to higher order problems, by considering all indexes as 2D.
- Obtaining possible y outcomes as matrix spectrum (10) and then calculating their probabilities by projection (11) of given distribution (point distribution (7) is a simplest example of such) to eigenvectors (13) is a powerful approach to estimation of y distribution under given condition. We can expect this approach to show good performance for data drawn from a wide range of probability distributions, especially for distributions that are not normal. The reason is because the (10) is expressed in terms of probability states, what make the role of outliers much less important, compared to methods based on L^2 norm, particularly least squares. For example, this approach can be applied to distributions where only first moment of y is finite, while the L^2 norm approaches require second moment of y to be finite, what make them inapplicable to distributions with infinite standard deviation. We expect the (10) approach can be a good foundation for construction of Robust Statistics[9].

[1] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez, “Solving the multiple instance problem with axis-parallel rectangles,” *Artificial intelligence* **89**, 31–71 (1997).

- [2] Jun Yang, *Review of multi-instance learning and its applications*, Tech. Rep. (Tech. Rep, 2005).
- [3] Zhi-Hua Zhou, “Multi-Instance Learning: A Survey,” Department of Computer Science & Technology, Nanjing
- [4] Zoltán Szabó, Arthur Gretton, Barnabás Póczos, and Bharath Sriperumbudur, “Learning theory for distribution regression,” arXiv preprint arXiv:1411.2066 (2014).
- [5] Vladislav Gennadievich Malyshkin and Ray Bakhramov, “Mathematical Foundations of Real-time Equity Trading. Liquidity Deficit and Market Dynamics. Automated Trading Machines. <http://arxiv.org/abs/1510.05510>,” ArXiv e-prints (2015), arXiv:1510.05510 [q-fin.CP].
- [6] Vladislav Gennadievich Malyshkin, “Radon–Nikodym approximation in application to image analysis. <http://arxiv.org/abs/1511.01887>,” ArXiv e-prints (2015), arXiv:1511.01887 [cs.CV].
- [7] Vladislav Gennadievich Malyshkin, “Multiple–Instance Learning: Christoffel Function Approach to Distribution Regression Problem,” ArXiv e-prints (2015), arXiv:1511.07085 [cs.LG].
- [8] Vladislav Gennadievich Malyshkin, (2014), the code for polynomials calculation, <http://www.ioffe.ru/LNEPS/malyshkin/code.html>.
- [9] Peter J Huber, *Robust statistics* (Springer, 2011).